Language Identification for Generating GIS Data Used in Mapping Linguistic Features of the World's Languages

Ren Wu[1], Hideyuki Inui[2], Manabu Sugii[3] and Hiroshi Matsuno[1]
[1]Graduate School of Science and Engineering, Yamaguchi University
[2]Faculty of Humanities, Yamaguchi University
[3]Media and Information Technology Center, Yamaguchi University
[1,2,3]Yoshida 1677-1, Yamaguchi City, Yamaguchi, 753-8512, Japan
E-mail : [1]{wu, matsuno}@ib.sci.yamaguchi-u.ac.jp, [2,3]{inui, manabu}@yamaguchi-u.ac.jp

Abstract: For the purpose of mapping linguistic features of the world's languages, it is necessary to identify the languages in Yamamoto-Data and SilGIS-Data. In this paper, firstly we point out the problems in the primary method that uses language name(s) for language identification. Secondly, noticing that the world's languages are classified and grouped into languages family trees, we propose an improved method for the language classification. Finally, we give our experimental results.

1. Introduction

GIS is used widely in various fields of studies, and recently GIS used in linguistic approaches have also become popular. Our objective is to analyze word order of the world's languages from the typological aspect using GIS. Firstly, we aim at mapping word order of the world's languages [1].

Generally in language studies using GIS data, the spatial data and the attribute data are necessary and important. The first one shows the positions where the languages are spoken, and the second one shows various linguistic features (word order is one of them) of the languages. Yamamoto surveyed of 2,932 languages and gave a set of word order data (called **Yamamoto-Data** hereafter) in [2]. On the other hand, WLMS (The World Language Mapping System, hereafter called **SilGIS-Data**) [3], a data set for GIS, includes the spatial data that is necessary in word order mapping process.

Based on Yamamoto-Data, an approach to generate essential GIS data from SilGIS-Data was proposed [1], and the problem firstly being settled was the identification of the languages between Yamamoto-Data and SilGIS-Data. The proposed method (called the primary method hereafter) tried to find the corresponding language in SilGIS-Data for every language of Yamamoto-Data. We succeeded in finding the corresponding languages for about 66% languages of Yamamoto-Data. However there are still many languages undecided [1]. One more problem is that we had no methods to distinguish one language from another language that have the same name.

Languages of the world have been classified and grouped into many of tree structures (called **Family-Tree** hereafter). Those Family-Trees are not invariable. Any one language of Yamamoto-Data and SilGIS-Data is expressed as a leaf node in Family-Tree. In the primary method, only the node name corresponding to the name of the language was taken into account, but the node position in Family-Tree (i.e., the path from the root to itself) was not. The node position seems to be a factor to identify the languages. Therefore, based on Family-Trees, the unidentified languages might be identified.

This paper is organized as the following. We firstly describe the formats of Yamamoto-Data and SilGIS-Data. After introducing the primary method and pointing out its problem, we propose a new improved method. Finally, we show the results by our improved method.

2. The Necessity of Language Identification and the Data Formats

Here we introduce data formats of Yamamoto-Data and SilGIS-Data.

Table 1 shows an example of Yamamoto-Data. In Table 1, each record indicates exactly one language. For each language it gives the Language Name (denoted by LN), Classification Code (denoted by CC) and linguistic features (*Clase* express its word order which is one of linguistic features) etc. The classification scheme of CC used here is basically that of *Ethnologue 12*th [4] and the code scheme is defined by Yamamoto on his own right. Here a unique three-letter Language Code (denoted by LC) is not given.

SilGIS-Data contains shapefiles [5] that stores geographic features of the world's languages. Geographic features involve spatial information that is the geometric location as well as attribute information. Attribute information is stored in dBASE tables, which are file-based tables, and can be joined to a shapefile's geometric location. Table 2 shows an example of the dBASE table which contains that all-inclusive attribute data of SilGIS-Data. Hereafter, SilGIS-Data will indicate this attribute data contained in the dBASE table.

SilGIS-Data is compliant with *Ethnologue 15*th [6]. In Table 2, for each record it gives a unique three-letter Language Code (denoted by LC) that is the codes of ISO/DIS 639_3 [7], Primary language Name (denoted by PN) [6], Alternate Name(S) (denoted by ANS) [6] and FIPS10-4 contry code (denoted by $FIPS$) [8] etc. Languages are often spoken across national boundaries. In Table 2, a string $LC{-}FIPS$ combined with LC and $FIPS$ is the record identifier, thus two or more records may probably indicate a same language.

In Tables 1 and 2, No is appended for convenience and is in the order of LN and $LC{-}FIPS$ respectively. LN and PN both consist of single word or multiple words connected with a separation character and express a language name. Many languages are spoken in more than one country and known by more than one name [6]. Different scholars probably prefer different names as main names. In ANS, multiple language names are concatenated with commas. Also,

Table 1. Yamamoto-Data

No	LC	LN	CC	Clause $\cdots$
151		ARANDA,WESTERN	ALPAU	SOV $\cdots$
217		BAI	NCVNUSSB	SVO/svo $\cdots$
218		BAI	STTLLM	SVO $\cdots$
1038		JAPANESE	JAJ	SOV $\cdots$
1733		MONPA,CENTRAL	STTOEE	SOV $\cdots$
2099		POMO:SOUTHEAST	HONP	SOV $\cdots$

Table 2. SilGIS-Data

No	LC	$FIPS$	PN	ANS	$\cdots$
532	are	AS	ARRARNTA,WESTERN	Aranda, Arunta	$\cdots$
3908	jpn	JA	JAPANESE		$\cdots$
3920	jpn	US	JAPANESE		$\cdots$
7558	pom	US	POMO, SOUTHEASTERN	Lower Lake Pomo	$\cdots$
9389	tsj	CH	TSHANGLA	Sangla, $\cdots$, Central Monpa	$\cdots$

there are many languages that appear as different names in Yamamoto-Data and SilGIS-Data, although these two data sets created from the same *Ethnologue* but different edition.

If Tables 1 and 2 have the same identifier, we can map word order by joining Table 1 with Table 2 [1]. But LC in Yamamoto-Data is not given currently, in addition LN and PN of the same language may be not the same. Thus we have to firstly identify languages in both data, so that LC of Yamamoto-Data can be found out and gets to map word order.

The following definition gives data formats of Yamamoto-Data and SilGIS-Data for the later use in this paper.

[Definition 1]

(1) Yamamoto-Data is expressed by $DATA^Y = \{R_1^y, R_2^y, \ldots\}$. Each element of $DATA^Y$ is 4-tuple $R_i^y = (LC_i^y, LN_i^y, CC_i^y, RE_i^y)$ (i corresponds to No of Table 1), where (i) LC_i^y is three-letter code satisfying $LC_i^y{=}\phi$; (ii) $LN_i^y{=}left_i^y{:}right_i^y$ expresses language name (LN). Here list $left_i^y{=}(W_{i1}^y, W_{i2}^y, \ldots)$; $right_i^y$ has the same format as $left_i^y$ and may be $right_i^y{=}\phi$; (iii) CC_i^y is the classification code; (iv) RE_i^y represents the set of remaining elements of R_i^y.

(2) SilGIS-Data is expressed by $DATA^{E15}{=}\{R_1^{e15}, R_2^{e15}, \ldots\}$. Each element of $DATA^{E15}$ is 4-tuple $R_i^{e15} = (PN_i^{e15}, LC_i^{e15}, ANS_i^{e15}, RE_i^{e15})$ (i corresponds to No of Table 2), where (i) $PN_i^{e15}{=}left_i^{e15}$ is a list with the same format as $left_i^y$; (ii) LC_i^{e15} is three-letter code and $LC_i^{e15}{\neq}\phi$; (iii) $ANS_i^{e15}{=}(AN_1^{e15}, AN_2^{e15}, \ldots)$, and these alternate name AN_j^{e15} are concatenated with "," and are lists with the same format as PN_i^{e15}. □

The words in LN^y and PN^{e15} both can be divided into two types, expressing main information (e.g., ARANDA, JAPANESE) and extra information (e.g., WESTERN, CENTRAL) of a language. We denote the set of extra information by EI for the later use.

3. Two Methods for Language Identification

In this section, we are going to describe two methods for language identification.

3.1 The primary method

In primary method, we have two processes as stated in the following A and B.

A. Deleting duplicatively named languages of $DATA^Y$

For either of Yamamoto-Data or SilGIS-Data, it is probably that two different languages are expressed by the same LN_i^y or PN_j^{e15}. Such languages are called Duplicatively Named languages (called **DN-languages** hereafter). Because we have no method to determine which language is our target, the records of DN-languages are to be deleted in our processing. We focus on the following types of DN-languages.

[Definition 2]

Type 1: Let $DATA\_DUP1^Y$ be a set of records of $DATA^Y$ such that for each $R_i^y{\in}DATA\_DUP1^Y$ there exists at least a record $R_j^y{\in}DATA\_DUP1^Y$ satisfying $PN_j^y{=}LN_i^y$ ($i{\neq}j$). The languages indicated by the records of $DATA\_DUP1^Y$ are called *Type 1* languages.

Type 2: Let $DATA\_DUP^{E15}$ be a set of records of $DATA^{E15}$ such that for each $R_i^{e15}{\in}DATA\_DUP^{E15}$ there exists at lease a record $R_j^{e15}{\in}DATA\_DUP^{E15}$ satisfying $PN_j^{e15}{=}PN_i^{e15}$ ($i{\neq}j$). And let $DATA\_DUP2^Y$ be a set of records of $DATA^Y$ such that for each $R_i^y{\in}DATA\_DUP2^Y$ (i) $R_i^y \notin DATA\_DUP1^Y$; (ii) there exists $R_j^{e15} \in DATA\_DUP^{E15}$ satisfying $PN_i^y = PN_j^{e15}$. The languages indicated by the records of $DATA\_DUP2^Y$ are called *Type 2* languages. □

Deleting the records of DN-Languages, we obtain the remaining records $DATA\_P^Y$, i.e., $DATA\_P^Y{=}DATA^Y{-}DATA\_DUP1^Y{-}DATA\_DUP2^Y$.

B. Searching the same language names from SilGIS-Data

For each word of LN_i^y included in $R_i^y{\in}DATA\_P^Y$, we search the same word in SilGIS-Data in the order of PN and ANS. We use the notions, $\leftrightarrow$, $\sim$ and $\nsim$, to express our search results as defined in the following.

[Definition 3] Let LN^y and PN^{e15} be two entries of $R^y{\in}DATA\_P^Y$ and $R^{e15}{\in}DATA^{E15}$, respectively.

(1) Let W_{LN} and W_{PN} be the sets of words of LN^y and PN^{e15}. (i) If $|W_{LN}|{=}|W_{LN}{\cap}W_{PN}|$, i.e., all the words of LN^y are included in PN^{e15}, then we say LN^y perfectly matches with PN^{e15} and denote it by $LN^y{\leftrightarrow}PN^{e15}$; (ii) if $|W_{LN}| > |W_{LN}{\cap}W_{PN}|{\geq}1$ and $W_{LN}{\cap}W_{PN}{\not\subset}EI$, i.e., there is at lease one word that is included in PN^{e15} and not included in EI, then we say LN^y partly matches with PN^{e15} and denote it by $LN^y{\sim}PN^{e15}$; (iii) if it is not the case of (i) or (ii), then we say LN^y does not match with PN^{e15} and denote it by $LN^y{\nsim}PN^{e15}$.

(2) Let ANS^{e15} be an entry of R^{e15}. Consider ANS^{e15} as W_{PN}, we similarly give the followings as (1): (i) if

$right^y$ of LN^y satisfies $right^y{=}\phi$ and the condition of (1)-(i) holds, then $LN^y{\leftrightarrow}ANS^{e15}$; (ii) if the condition of (1)-(ii) holds, $LN^y{\sim}ANS^{e15}$; (iii) if the condition (1)-(iii) holds, then $LN^y{\nsim}ANS^{e15}$. □

Since a language is indicated by a record (R^y) of $DATA^Y$, we classify the languages into the following three levels according to our search results.

level-1: R^y's language is level-1, if there is a R^{e15} such that $LN^y{\leftrightarrow}PN^{e15}$ or $LN^y{\leftrightarrow}ANS^{e15}$ holds;

level-2: R^y's language is level-2, if there is a R^{e15} such that $LN^y{\sim}PN^{e15}$ or $LN^y{\sim}ANS^{e15}$ holds;

level-3: R^y's language is level-3, if it is neither level-1 nor level-2.

We determine $LC^y{=}LC^{e15}$ for each record R^y of $DATA^Y$, which satisfies $LN^y{\leftrightarrow}PN^{e15}$.

3.2 An improved method

A. Language family tree

In the primary method, the languages of level-2 and level-3 are not resolved yet as well as DN-languages. It is only based on language names (primary language name, alternate name(s)) without considering factors of language classification.

Languages of the world are classified linguistically into many of language families (the genealogical classification may not include all), and each language family appears as a Family-Tree. In addition, the Family-Tree defined by different linguists maybe different from each other. Furthermore, languages in a Family-Tree have two cases: (1) one is that a language is located at the leaf of the Family-Tree; (2) another is that a language is subdivided into another single or multiple languages. Hereafter, the former is called Language (indicated by elliptically-shaped nodes in Figure 1), and the latter is called Group (indicated by box-shaped nodes in Figure 1).

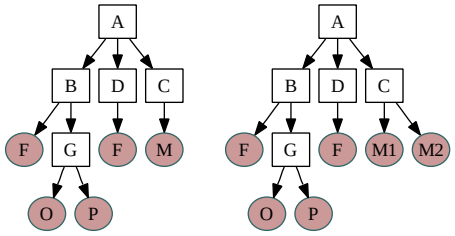
Actually, every record R_i^y or R_j^{e15} indicates a Language. In Figure 1, two "F"s have different paths "A-B" and "A-D", and show a case of DN-languages. Obviously, the primary method cannot give a resolution for these LN_i^y included in $R_i^y{\in}DATA^Y$, and so we have to delete them from processing data.

"M" in Figure 1 shows a case of those Languages that is a language before and subdivided into multiple Languages "M1" and "M2" (called **SD-languages** hereafter). Similarly, if there exist any LN_i^y included $R_i^y{\in}DATA^Y$ as the same as SD-languages, LN_i^y will never be identified as a level-1 language by the primary method.

Therefore, besides the language name (i.e., the node name in a Family-tree), if the factor of language classification (i.e., the path from root in a Family-tree) is considered, perhaps the languages such as the cases of "F" or "M" will possibly get to be identified.

B. Language family index data

Language family index data, corresponding to Yamamoto-Data (called Yamamoto-FI-Data hereafter) and SilGIS-Data (because SilGIS-Data is compliant with *Ethnologue 15th*, we call it E15-FI-Data hereafter) respectively which can be used in constructing Family-Trees, are also given by Yamamoto [2] and website [9].



(a) Family-Tree A (b) Family-Tree B

Figure 1. Examples of DN-languages and SD-languages

Table 3. Yamamoto-FI-Data

CL	CC	Clause	$\cdots$
Japanese	JA		$\cdots$
*Japanese	JAJ	SOV	$\cdots$
*Ryukyuan	JAR	SOV	$\cdots$
**Amami-Okinawan	JARA		$\cdots$
***Northern Amami-Okinawan	JARAN		$\cdots$
***Southern Amami-Okinawan	JARAS		$\cdots$
**Sakishima	JARS		$\cdots$

Yamamoto-FI-Data is mainly compliant with the language family index of *Ethnologue 12th*. Yamamoto-FI-Data contains a part of language family index of *Ethnologue 12th* that are associated with the languages of Yamamoto-Data.

E15-FI-Data is language family index of *Ethnologue 15th*. The revision of *Ethnologue* is done regularly. During the period time of the revision from *Ethnologue 12th* to *Ethnologue 15th*, the number of languages grew from 4,493 to 6,809 [6]. This means that the case of "M" as in Figure 1 may have occurred.

Table 3 shows an example of Yamamoto-FI-Data. In Table 3, CLassification (denoted by CL) is a family name (there exists no vanward "*") or group name (or subgroup name) where LN_i^y belongs to, and vanward "*" expresses ranks. In Yamamoto-FI-Data, LN_i^y itself does not appear in it, but its Classification Code (denoted by CC) that is based on the same code schema as Yamamoto-Data is included in it and is unique. That, LN_i^y can be known that which family and group (or subgroup) it is classified to through CC_i^y.

Language family trees of Yamamoto-FI-Data (denoted by LFT^Y) and E15-FI-Data (denoted by LFT^{E15}) can be constructed from Yamamoto-FI-Data and E15-FI-Data (we succeeded in taking E15-FI-Data from the website [9] automatically).

LFT^Y and LFT^{E15} are defined as the following.

[Definition 4] Let $LFT^Y{=}(N^Y, E^Y)$ and $LFT^{E15}{=}(N^{E15}, E^{E15})$ be the language family trees corresponding to Yamamoto-FI-Data and E15-FI-Data respectively, where,

(1) N^Y and N^{E15} are node sets of LFT^Y and LFT^{E15}, respectively. Each $n_i^y{\in}N^Y$ consists of LN_i^y and each $n_i^{e15}{\in}N^{E15}$ consists of PN_i^{e15} and ANS_i^{e15}. Each node represents a *Family*, or a *Group*, or a *Language*. Especially, a leaf node represents a *Language*;

(2) E^Y and E^{E15} are directed edge sets of LFT^Y and

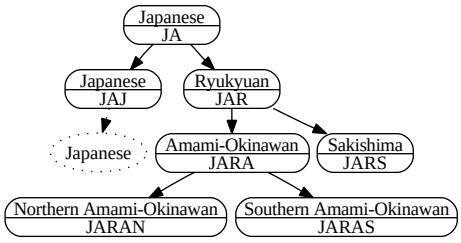


Figure 2. "Japanese" Family-Tree in LFT^Y, in which the position of LN^y_{1038} ="Japanese" can be settled by CC^y_{1038}="JAJ" and added as a leaf node.

LFT^{E15}, respectively. Each $e^y_i \in E^Y$ or $e^{e15}_i \in E^{E15}$ connects from a *Family* to a *Group* / *Language*, or from a *Group* to its *Subgroup* / a *Language*. □

Figure 2 shows "Japanese" language family's family tree that is an example of LFT^Y.

C. Searching the same path and node

Our new method is to judge if two languages are the same. Since each node n^y_i (n^{e15}_i) of LFT^Y (LFT^{E15}) consists of LN^y_i of R^y_i (PN^{e15}_i and ANS^{e15}_i of R^{e15}_i), we say the two language indicated by R^y_i and R^{e15}_i are the same if the following conditions are satisfied:

Condition 1: R^y_i's language is level-1; and further

Condition 2: Two directed paths, from the roots to n^y_i in LFT^Y and to n^{e15}_i in LFT^{E15}, are exactly the same except themself.

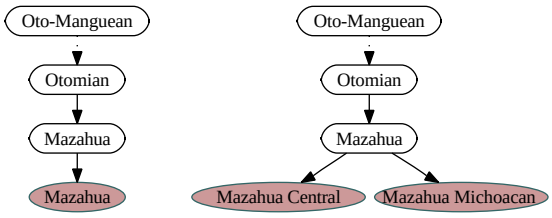
4. Experimental Results

We have done experiments by using the primary method and the improved method to identify languages between Yamamoto-Data and SilGIS-Data. Yamamoto-Data and SilGIS-Data include respectively 2,932 and 10,512 records.

At first, we used the primary method to do the experiment and obtained the results: (i) 87 DN-languages were found in Yamamoto-Data and deleted in our processing; (ii) searching the same names in SilGIS-Data for each record of Yamamoto-Data, we got 2,173 level-1 languages, 672 level-2 and level-3 languages. For the 2,173 level-1 languages, we determine their three-letter language codes (LC).

Although 2,173 languages were identified, they were not confirmed. Therefore, we apply the improved method to all 2,932 languages of Yamamoto-Data and continuously do the experiment. The results are: (iii) including 11 SD-languages, 972 languages have exactly the same paths between LFT^Y and LFT^{E15}, and can be correctly determined (Figure 3 shows an example of SD-languages identified newly in our improved method); (iv) the remaining 1,960 languages, including those level-2, level-3 and also the level-1 languages with different paths, were not determined.

We have to say, any language identified by the primary method can be assigned with three-letter code, but not all identified languages by the improved method can do so. SD-languages are those identified ones but cannot be assigned three-letter codes languages.



(a) LFT^Y (b) LFT^{E15}

Figure 3. An example of experimental results

5. Concluding Remarks

We have given a formal description of Yamamoto-Data and SilGIS-Data, briefly introduced a primary method for language identification and proposed an improved method based on the languages family trees. Experiments have been done by using the improved method in order to evaluate its usefulness.

Our experimental results show that: (1) the improved method can correctly identify languages between Yamamoto-Data and SilGIS-Data; (2) this method seems to be too rigid to identify languages, since it cannot determine those same languages having the same name in Yamamoto-Data and SilGIS-Data.

As the future work, we need to investigate the changes of language family index data from *Ethnologue* 12th to *Ethnologue* 15th as well as Yamamoto-FI-Data, and further improve our identification method.

6. Acknowledgement

The authors would like to thank Professor Hideki Yamamoto, Hirosaki University, Japan, for kindly providing us Yamamoto-Data and answering our questions about Yamamoto-Data during this work.

References

[1] R.Wu, H. Inui, M. Sugii and H. Matsuno, "On generating GIS data for language studies:a method of mapping of language characteristic from *Ethnologue* GIS Data," *IPSJ SIG Computers and the Humanities Symposium(2007)*, pp.253-258, Dec 2007.

[2] Hideki Yamamoto, *Survey and Historical Study of Geographical and Genealogical Distribution of Word Order around the World*, Keisuisha Co. Ltd, Dec 2003.

[3] http://www.gmi.org/wlms/

[4] Barbara F. Grimes (ed.), *Ethnologue: Languages of the world, 12th ed.*, Dallas: Summer Institute of Linguistics, 1992.

[5] http://www.esrij.com/support/arcview3/material/shape/shapefile.pdf

[6] Gordon, R.G. (ed.), *Ethnologue: Languages of the World, 15th ed.*, Dallas: SIL International, 2005.

[7] http://www.sil.org/iso639-3/default.asp

[8] http://www.itl.nist.gov/fipspubs/fip10-4.htm

[9] http://www.ethnologue.com/family_index.asp